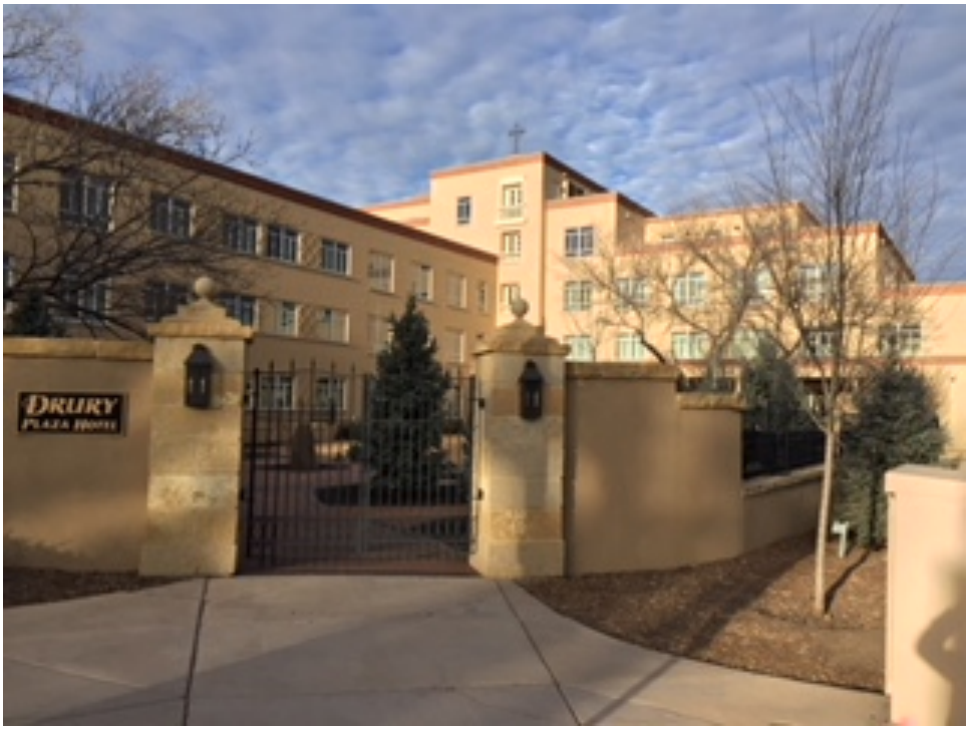
2018 Joint Research Conference on Statistics in Quality, Industry, and Technology

Theme: The Art and Science of Statistics



Santa Fe, New Mexico

June 11-14, 2018

Hosted by Los Alamos National Laboratory

LA-UR-18-24947

Program Information for the 2018 Joint Research Conference

June 11-14, 2018

Contents

Conference Honoree

Professor Max Morris

A brief, incomplete, and biased history of computer experiments

chair: Lisa Moore, Los Alamos National Laboratory

For several decades, statisticians and researchers from other professions have discussed how output generated from computer models can be usefully analyzed, sometimes along with other sources of data. In this talk, I'll review some of the efforts in the area broadly called "computer experiments". While I'll offer some examples of research and applications, the focus will be more on the big picture of how this work fits into modern statistics than with detailed technical descriptions. As we rely more and more on formal computational models to express what we know, analysis aimed at understanding and predicting what we don't know will require continued development of methodology that takes advantage of these models.

Plenary Speakers

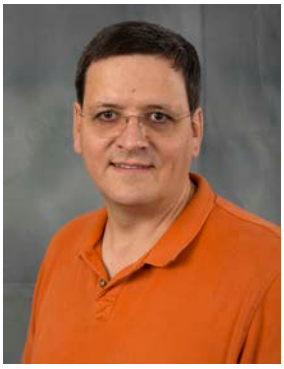
Derek Bingham

Computer model calibration: Applications from the national labs and a protein network

chair: Earl Lawerence, Los Alamos National Laboratory

The use of computational models to explore engineering and scientific processes is now commonplace. Computer model calibration uses outputs from a simulator and field data to build a predictive model for the physical system and to estimate unknown inputs. In this talk, I will touch on two main topics: (i) some recent work on the design and analysis of computer experiments for model calibration that is inspired by applications funded by the Department of Energy (DOE); and (ii) model calibration for a protein signalling network. In the first case, I will discuss collaborative research that was done with scientists at the national labs and through the DOE Predictive Science Academic Alliance Program. These include applications in cosmology, radiative shock hydrodynamics and radiation transport. In the latter case, a computational model for a protein signalling network is studied. Using experimental data helps constrain such networks and provide further insight into the physical system. In this talk, an approach for calibrating a protein signaling network model is presented. The data structure of interest is a complex network with non-linear transfer functions between nodes. The data include observations under different experimental stimuli, but with incomplete sampling of network nodes. We cast this application as a computer model calibration problem and develop a new sequential MCMC approach for exploring the network structure and quantifying uncertainty.

Scott Vander Wiel



Modeling stress-strain fields in polycrystalline materials-statistical art and science

chair: Brian Weaver, Los Alamos National Laboratory

Polycrystalline materials fail under shock-loading when stresses build up at weak spots and nucleate voids that grow and coalesce into macro-scale cracks and tears. Quantita-

tive understanding of these inherently statistical processes is needed to assure reliable performance of materials in extreme environments. I will describe collaborative statistical modeling that seeks to extract phenomenological patterns in simulated micro-scale stress-strain fields. Our eventual goal is to translate discoveries into continuum-scale simulations with porosity effects that are physically well-grounded.

Metals and other polycrystalline materials consist of many single-crystal grains of various sizes and shapes. Atoms within a grain are arranged in a regular lattice. Loading forces applied to an aggregate volume cause grains to morph along the lattice slip systems, resulting in spatially heterogeneous stresses and strains, especially at grain boundaries.

I will describe three statistical representations of stress-strain fields in an assembly of simulated Tantalum. First, a Gauss-Markov random field model on 1M computational elements is designed to capture stress effects that tend to become more extreme near the surface of grains and across grain boundaries. Next, two different regression models are fit to reduced data sets corresponding to grain boundary centroids. Orientations of crystal lattices are used as predictors in these models and present special challenges because the point group structure of crystal orientations imposes symmetries on the regression problem. One model regresses on (hyper-) spherical harmonic bases. The other is a Gaussian process fit utilizing an orientation distance metric.

The talk will highlight idiosyncrasies of building statistical models for polycrystalline materials and discuss some of the art of collaborating to solve problems in this domain. This talk is based on joint work with Peter Marcy, Curt Bronkhorst, and Veronica Livescu.

Lunch Speakers

Francesca Samsel

Color in scientific visualization

This presentation will discuss color in scientific visualization, drawing on a number of multidisciplinary collaborations between the arts, sciences, and visualization. Insights on color theory, the organizational potential of color, design principles, practical considerations, resources, and strategies, will be discussed. Methods and principles for selecting appropriate, effective, and affective palettes will be described. This research focuses on identifying artistic principles and expertise with potential to assist scientists in their scientific inquires. This work is based on collaborations with computational teams at Los Alamos National Laboratory, the Texas Advanced Computing Center, and the University of Minnesota Interactive Visualization Laboratory. Further information may be found at SciVisColor.org.

Jeff Hooper

Quality and Productivity Research Conference - 35 years

Xinwei Deng

Spring Research Conference - 25 years

Karen Kafadar

The critical role of statistics in development and validation of forensic methods

Statistics has played a key role in the development and validation of forensic methods, as well as in the inferences (conclusions) obtained from forensic evidence. Further, statisticians have been important contributors to many areas of science, such as chemistry (chemometrics), biology (genomics), medicine (clinical trials), and agriculture (crop yield), leading to valuable advances that extend to multiple fields (spectral analysis, penalized regression, sequential analysis, experimental design). Yet the involvement of statistics specifically in forensic science has not been well established, given the value it has demonstrated thus far (e.g., assessment of bullet lead evidence; significance of findings in the U.S. anthrax investigations, reliability of eyewitness identification) and its contributions to advances in other branches of science. In this talk, we first discuss three past cases in which statistics played a vital role in the findings. We then suggest ways in which statisticians can participate with the ultimate goal of strengthening forensic evidence to achieve its mission in achieving low error rates (false positives and false negatives), thereby raising the level of confidence in the forensic system.

Sponsors

We are grateful to the sponsors of JRC 2018:

- Los Alamos National Laboratory
  - Statistical Sciences group, CCS-6
  - Computer, Computational, and Statistical Statistical Sciences division
  - Associate Directorate for Theory, Simulation, and Computing
  - Principle Associate Directorate for Science Technology, and Engineering
  - LANL Protocol
- American Statistical Association
- Quality and Productivity section of the American Statistical Association

- Section on Physical and Engineering Sciences of the American Statistical Association
- Procter & Gamble
- SAS, JMP Division
- Trilobyte
- Minitab
- Statease
- Springer

Committees

Conference Organizers:

- Joanne Wendelberger, Los Alamos National Laboratory
- Brian Weaver, Los Alamos National Laboratory
- Xinwei Deng, Virginia Tech
- Devin Francom, Los Alamos National Laboratory

Organizing Committee:

- C. Devon Lin, Queens University
- Matt Pratola, The Ohio State University
- Anne Hansen-Musakwa, Tuft & Needle
- Shane Reese, BYU
- Maria Weese, Miami University

Short Course

Bridging Statistics and Data Science

Instructor: Ming Li, Amazon
Instructor: Hui Lin, Dow DuPont

With the recent big data revolution, enterprises ranging from FORTUNE 500 to startups across the US are using Data Science to bring valuable business insight from

all the data collected. Statisticians are great data scientist candidates, but there are relatively few data scientists with statistics education background. This course aims to bridge the gap between Statistics and Data Science. Data science is a combination of science and art with data as the foundation. We will cover both the science part and the art part (such as data science project flow, general pitfalls in data science projects, and soft skills to communicate with business stakeholders effectively). The course will be hands-on and we will use the Databricks community edition cloud platform and R-Studio to illustrate programming, big data platform usage (such as Spark) and standard machine learning algorithms.

Invited Program

Modern Design of Experiments

Organizer: C. Devon Lin, Queen's University
Chair: Shan Ba, Procter & Gamble

Analysis-of-marginal-Tail-Means - a new method for robust parameter optimization

Jeff Wu, Georgia Tech.

This paper presents a novel method, called Analysis-of- marginal-Tail- Means (ATM), for parameter optimization over a large, discrete design space. The key advantage of ATM is that it offers effective and robust optimization performance for both smooth and rugged response surfaces, using only a small number of function evaluations. This method can therefore tackle a wide range of problems, particularly in applications where the performance metric to optimize is "black-box" and expensive to evaluate. The ATM framework unifies two parameter optimization methods in the literature: the Analysis-of- marginal-Means (AM) approach (Taguchi, 1986), and the Pick-the- Winner (PW) approach (Wu et al., 1990). We show that, by providing a continuum between AM and PW via the novel idea of marginal tail means, the proposed method offers a balance between three fundamental trade-offs. By adaptively tuning these trade-offs, ATM can then provide excellent optimization performance over a broad class of response surfaces using limited data. We illustrate the effectiveness of ATM using several numerical examples, and demonstrate how such a method can be used to solve two real-world engineering design problems. (based on joint work with Simon Mak.)

Restricted screening designs

Ryan Lekivetz, JMP Division of SAS

For some screening experiments, physical limitations on the system restrict not only the number of experiments, but also the number of factors that can be considered in any

experimental run. We discuss a way to create these designs through balanced arrays and balanced incomplete block designs. We will also present practical considerations that have been observed for these design situations.

Value of feedback and lookahead in optimal sequential Bayesian experimental design

Ryan Xun Huan, Sandia National Laboratory

Experimental data play a crucial role in developing and refining models of physical systems. However, some experiments produce more useful data than others, and well-chosen experiments can provide substantial resource savings. Optimal experimental design (OED) thus seeks to systematically quantify and maximize the value of experiments. The optimal design of a sequence of experiments needs to account for feedback between experiments as well as future consequences from each design decision. Common practice such as batch design (no feedback) and myopic design (not forward-looking) are thus suboptimal. We formulate the optimal sequential design problem using dynamic programming (DP), and employ a Bayesian framework that seeks to maximize expected information gain. The DP problem is solved using approximate value iteration, where design policies are iteratively refined via backward induction and regression using samples from exploration and exploitation. We illustrate the value of feedback and lookahead in a sensor placement application, and provide insights on where these benefits are most important.

Lessons Learned From Data Challenges and Challenging Data

Organizer: Anne Hansen-Musakwa, Tuft & Needle
Chair: Brian P. Weaver, Los Alamos National Laboratory

Mattresses & CPUs: How data-based approaches work regardless of corporate culture

Anne Hansen-Musakwa, Tuft & Needle

Regardless of corporate culture and structure, data driven methodologies are an efficient way to determine focus areas and answer key business questions. The author will provide examples of using applied statistics and data driven methodology in two different corporate cultures– one large and centralized, the other bootstrapped and flat. We show that in both settings, statistical techniques are important tools to ensure product quality and wonderful customer experiences.

The value of the CDC's influenza forecasting challenge: One participant's nuanced perspective

David Osthus, Los Alamos National Laboratory

The Centers for Disease Control and Prevention (CDC) has historically been tasked with influenza (flu) surveillance in the United States. Flu surveillance involves thousands of hospitals, clinics, and laboratories providing data to the CDC. The CDC then organizes, synthesizes, and ultimately disseminates the information to the public. In recent years, the CDC has become interested in flu forecasting, not just surveillance. To better understand flu forecasting capabilities and to improve their usefulness to public health decision-makers, the CDC organized the inaugural flu forecasting competition for the 2013-14 flu season, where participants provided forecasts to flu-relevant targets throughout the flu season. The CDC's flu forecasting challenge has occurred every year since then, growing in popularity from 13 participating models in 2013-14 to 30 participating models in 2017-18.

In this talk, I'll describe in detail the CDC's flu forecasting challenge through the lens of a participant (I participated in the 2016-17 and 2017-18 challenges). Topics to be discussed include the different incentive structures for the challenge host versus challenge participants, how interesting/messy real-world data relates to statistical modeling/forecasting choices, a survey of the different participating forecasting models, and the value of the flu forecasting challenge to all stakeholders.

Data competition hosting: Getting more than just a winner through strategic design and analysis

Christine Anderson-Cook, Los Alamos National Laboratory

Leveraging the depth and breadth of solutions possible from crowdsourcing can be a powerful accelerator to method development for high consequence problems. While participants competing in data science competitions have become quite popular and prevalent, their implementations are highly variable and can sometimes lead to selecting a winner whose solution does not closely matched to the real problem of interest. This talk outlines considerations when hosting a competition, such as (1) defining the precise problem, (2) identifying the target audience, (3) using design of experiment strategies to specifying data sets to include for testing for interpolation and extrapolation to new scenarios, (4) determining a robust and relevant scoring metric that appropriately orders the competitors to match study goals, and (5) developing an analysis to provide summaries beyond just a winner on the leaderboard. The methods will be illustrated using a current competition to evaluate algorithms capable of detecting, locating, characterizing radioactive materials, which is of particular interest to the Office of Proliferation Detection (OPD).

Uncertainty Quantification

Organizer: Matt Pratola, The Ohio State University
Chair: Derek Bingham, Simon Fraser University

Neural networks for flexible and fast emulation of computer experiments

Jared Huling, The Ohio State University

The use of Gaussian processes in the emulation of complex computer experiments has seen wide application in the literature. However, the increased scale and complexity of computer simulations and similar problems has resulted in the need for scalable and accurate emulators, as gaussian process emulators are difficult to compute for large experiments and can struggle in high dimensions. Deep neural networks provide an attractive alternative for emulation as they have been widely successful for prediction tasks with high complexity and are trivially scalable. There are several obstacles to the application of neural networks to computer experiments: they are not directly suitable for emulation and prediction intervals are not immediately available. To address these challenges, we propose a class of nonstationary covariance functions based on neural networks. The resulting emulator is highly flexible, can effectively handle high dimensional problems, and provides a natural framework for constructing prediction intervals. It is also computationally efficient for large scale applications, often adding little computational burden beyond simply training a neural network. We demonstrate its performance on simulation studies and an application in inverse kinematics.

Bayesian Gaussian process models for dimension reduction uncertainties

Peter Marcy, Los Alamos National Laboratory

It is often known that important directions within the input space of Gaussian process (GP) regression models do not align with the original coordinate directions. It can be the case that changes along unknown linear combinations of the inputs effect the biggest changes in the response variables. In this scenario, Bayesian estimation of the dimension reduction space and understanding the posterior distribution are both fraught with difficulties due to the manifold structure of the parameter space. In this talk I describe the challenges and present a natural parameterization that lends itself to a relatively clean interpretation of the dimension reduction uncertainty. The methodology is illustrated via a sensitivity analysis of a complex computer model.

Model calibration and validation with count data and generalized Gaussian process emulation

Michael Grosskopf, Los Alamos National Laboratory

The calibration and validation of computer models plays an important role in building

trust in their use for science and engineering. We present work motivated by research modeling radiation transport at the Center for Exascale Radiation Transport (CERT). There, high performance computing is used for simulating the flow of energy by radiation in physical systems. We use a generalization of the Kennedy-O'Hagan calibration framework to handle the count structure of our observed data, allowing closer fidelity between the generating process of the model and the experimental system. This enables model validation by comparison of the distribution predicted by the statistical model using the simulator and that observed in experiment, taking into account epistemic uncertainties. We present this work as the foundation of validation of the radiation transport model across a hierarchy of physical experiments.

Mary G. and Joseph Natrella Scholarship Award Session

Organizer: Will Guthrie, NIST
Chair: Scott Kowalski, Minitab

Monitoring stochastic textured surfaces

Anh Tuan Bui, Northwestern University

We develop statistical process control methods to detect local defects and global changes for manufactured products whose quality is reflected by surfaces with stochastic nature (e.g., textiles or material microstructures). We use generic supervised learning methods to characterize the stochastic behavior of ?normal? in-control samples of the textured surfaces. For local defects, we propose two spatial moving statistics for detecting local aberrations in the textured surfaces, based on the residuals of the supervised learning model applied to new samples. For global changes, we develop a monitoring statistic using likelihood-ratio principles to detect changes in the surface nature, relative to the in-control one. Our approaches can detect general surface defects or changes without requiring prior knowledge about them. We illustrate the approaches using simulated and real examples.

Engineering-driven data analytics for quality improvement

Xiaowei Yue, Georgia Institute of Technology

My research is focused on engineering-driven data analytics for quality improvement. I will show how engineering-driven data analytics can help to improve quality and productivity by two examples:

(1) High-dimension nonlinear data are commonly encountered in many complex engineering systems. It is an important and challenging topic to conduct feature extraction, monitoring, data mining and analytics based on these datasets. Conventional statistical or machine learning methods, such as mixed-effects model and tensor decomposition,

have several limitations including inability to separate detailed components, computational inefficiency, and high-dimensionality in data analysis. In this talk, we propose a novel method called a penalized mixed-effects decomposition (PMD) method, which can decompose the profiles into fixed effects, normal effects, defective effects, and noise. An accelerated proximal gradient (APG) based optimization algorithm has been proposed to realize efficient parameter estimation. The computational properties of this method are explored. Using surrogated data analysis and real case studies, we evaluated the performance of the proposed PMD method, and demonstrated the effectiveness and efficiency with applications to quality detection of Carbon Nanotubes buckypaper manufacturing process.

(2) Dimensional shape control of composite parts is vital for large-scale production and integration of composite materials in the aerospace industry. The current industry practice of shape control uses passive manual metrology. This has three major limitations: (i) low efficiency: it requires multiple trials and a long time to achieve the desired shape during the assembly; (ii) non-optimal: it can realize acceptable deviation reduction instead of optimal one; and (iii) experience-dependent: highly skilled engineers are required during the assembly process. We proposed an automated shape control system that can adjust composite parts to an optimal configuration in an effective and efficient manner. The objective is accomplished by (i) building a finite element analysis (FEA) platform, calibrated and validated by experimental data; (ii) developing a surrogate model with consideration of actuator uncertainty, part uncertainty, modeling uncertainty, and unquantified error to achieve accurate prediction, and then embedding the model into a feed-forward control algorithm; (iii) conducting multivariable optimization to determine the optimal actions of actuators. We show that the surrogate model considering uncertainties (SMU) achieves satisfactory prediction performance and that the automated optimal shape control system can significantly increase the productivity with improved dimensional quality.

Test Planning for Reliability

Organizer: Brian Weaver, Los Alamos National Laboratory
Chair: Caleb King, JMP Division of SAS

New developments on binomial demonstration test plans

Lu Lu, University of South Florida

Reliability demonstration testing (RDT) is an important reliability assurance activity to demonstrate the capability of a product to meet specified quality requirements. Binomial demonstration tests are often used for one-shot tests at the end of a fixed test duration. Zero-failure tests have been commonly used for their simplicity and minimum sample size requirements, but they have potential problems. This presentation discusses some recent developments for new binomial demonstration test plans that provide a broader and better set of options. The first category considers a Pareto front based RDT strategies based

on considering multiple risk and cost criteria. The method provides a structured process and graphical tools for selecting optimal test plans that simultaneously balance multiple criteria based on the consumer's and producer's risk levels, probability for passing a test, and the cost. The second category considers demonstration tests to meet multiple reliability requirements for either multiple time periods or failure modes. The third category considers a cost-driven strategy for selecting demonstration test plans based on a new comprehensive cost structure spanning the product lifetime including costs from the demonstration test, warranty claims, and reliability growth. The methods are illustrated through examples with guidance and insights offered for general applications.

Mutual information design criterion for sensitivity testing

Isaac Michaud, North Carolina State University

A sensitivity test is a sequence of experiments used to estimate the probability of failure or detonation for a given level of a stimulus variable. The choice of the stimulus levels in a test impacts the e?ciency of estimating these probabilities. Currently, the Neyer and 3pod methods, both based on sequential D-optimal designs for binary regression, are the best available methods for designing sensitivity testing. Unfortunately, there is little work demonstrating the difference between these methods when few experimental units are available to the experimenter. In this talk, a Bayesian method for designing sensitivity tests utilizing mutual information will be presented. The differences between the Neyer and 3pod methods will be explored as Bayesian experiments utilizing different prior uncertainties.

Challenges and new methods for designing reliability experiments

Laura Freeman, Institute for Defense Analysis

Reliability experiments are used to determine what factors drive product reliability, build robust products, and predict reliability under standard use conditions. Often, the reliability or lifetime data collected in these experiments tend to follow distinctly non-normal distributions and include censored observations, which occur when products do not fail within the allotted test time. Historical approaches to reliability experiments leverage classical experimental design techniques paired with an analysis that appropriately reflects the data. The experimental design should accommodate the skewed nature of the response and allow for censored observations. Current research in reliability experiments include optimal experimental design and new methods for evaluating designs. A core challenge is that ?optimal? designs depend on the complex balance of desired precision, power, effect size, model formulation, sample size, censoring rates, and design type. Another challenge is restricted randomization design, which make execution easier, but complicates the appropriate analysis. This paper covers current research in designing reliability experiments. New contributions by the authors in power analysis and restricted randomization are highlighted. We also discuss complex system reliability

and how research in product design might translate to systems with multiple functions and multiple use profiles, through the use of experimental design methods.

Astrostatistics Interest Group Invited Session

Organizer: Jessi Cisewski, Yale
Chair: Jogesh Babu, Penn State University

A new spectrum estimation technique for asteroseismology

Gwendolyn Eadie, University of Washington

Asteroseismology is the study of stellar oscillations. By observing a star's change in brightness over time and studying periodic components in the power spectrum, inference can be made about a star's internal processes, and theories about stellar evolution can be tested. Of particular interest to astronomers are the frequency of the maximum oscillatory power and the spacing between consecutive modes in the power spectrum. Recently, thanks to the Kepler spacecraft, there has been an increase in the amount of data for bright stars which have oscillation patterns similar to our Sun, such as red giants. Kepler was launched in 2009 with the primary mission of detecting Earth-like planets around other stars in our region of the Milky Way Galaxy. The spacecraft collected data on the brightness of stars over time with the goal of detecting transiting planets. The Kepler data set also gives astronomers the opportunity to study the asteroseismology of thousands of stars. Moreover, many stars have records exceeding four years in length (although some gaps in the data do exist).

In this talk, I will describe a new time series analysis method as it applies to Kepler data for red giant stars. The method we apply (Springford, 2017) combines the Thomson (1982) multitaper approach with the Lomb-Scargle periodogram. The multitaper-Lomb-Scargle spectrum estimate both greatly reduces spectral leakage and applies to data unevenly sampled in time. This method has the potential to uncover oscillations not found by traditional methods while also measuring known oscillations with more confidence.

Hunting for exoplanets around active stars

David Stenning, Imperial College London

The detection and characterization of exoplanets?planets that orbit stars other than the Sun?is one of the most active areas of research in modern astronomy. Many exoplanets are discovered using the radial velocity technique, which involves detecting the Doppler shift in a star?s spectral lines resulting from the gravitational effects of an orbiting planet. A challenge to this approach is that measured radial velocity signals are often corrupted by stellar activity such as spots rotating across the star?s surface. A principled method for recovering the underlying planetary radial velocity signal was

proposed by Rajpaul et al. (2015), which uses dependent Gaussian processes to jointly model the corrupted radial velocity signal and proxies for stellar activity. Our approach extends Rajpaul et al. (2015) by (i) incorporating science- and data-driven dimension reduction techniques to extract more informative stellar activity proxies, and (ii) introducing a model comparison procedure to select the best model for the stellar activity proxies at hand from a larger class of models. Our methodology is tested on synthetic data generated by the SOAP 2.0 code (Dumusque et al. 2014), and initial results show substantially improved statistical power for planet detection compared to using existing models from the literature.

Disentangling astronomical sources with spatial, spectral and temporal X-ray data

Luis Campos, Harvard

Spatially unresolved astronomical sources are difficult to study because the spatial domain is the primary way we disentangle sources. This is true even of nearly unresolved X-ray point sources where spectral extraction regions are defined manually to allocate source events. We present new methodology that uses spatial information (event locations and point spread functions), spectral information (photon energies), as well as event time information to aid in disentangling multiple sources. These Bayesian methods allow us to infer source parameters of interest, including the source location, relative brightness, and background intensity, all while systematically accounting for the uncertainty in those parameters. An added benefit of Bayesian modeling is that because the photons are probabilistically assigned to sources, further analyses can now account for this uncertainty. We show through simulation that when the event time and spectral information is meaningfully different between sources we are able to improve estimation of source specific parameters as well as properly allocate photons to sources. We find this holds even when the sources are nearly unresolved. We use this new methodology to study a pre-main sequence binary in a cluster of Young Stellar Objects in the Orion nebula, HBC 515.

Celebrating 50 Years of the Journal of Quality Technology

Organizer: Bianca Maria Colosimo, Politecnico Di Milano
Chair: Bianca Maria Colosimo, Politecnico Di Milano

The 50th anniversary of the Journal of Quality Technology

Douglas Montgomery, Arizona State University

In celebration of the 50th anniversary of the Journal of Quality Technology, this presentation takes a historical look at the articles that have appeared in the journal. Some of the key articles across the decades are highlighted. Some historical analysis on those

that have appeared is provided. This analysis illustrates some trends in the quality technology field and gives some insight about its future focus. This presentation is based on the article "50 years of the Journal of Quality Technology (with discussion)", by W.A. Jensen, D.C. Montgomery, F. Tsung, and G.G. Vining, Journal of Quality Technology, Vol. 50, No. 1, 2018, pp. 2-16.

Estimating a service-life distribution based on production counts and a failure database

Michael Hamada, Los Alamos National Laboratory

We assess the service life distribution of a product, where service life is the product lifetime after installation (not manufacture). The available data consists of production counts and a database providing information on failing units that is imperfect (i.e., missing information). For the most part, the database tracked repairs of an expensive repairable product warranted against repairs. The failure of interest here was relatively rare and driven by a different mode/mechanism than ordinary repair events. We develop data models for the service on a standard parametric lifetime distribution as well as a related limited failure population. We use the models to develop expressions for the likelihood of the available data that properly accounts for information missing in the database. Using a Bayesian approach, we investigate various characteristics of the service life distribution. We also discuss how a simulation experiment is used to validate the correctness of the custom software and the behavior of the statistical methodology before using its results in the application. This talk is based on joint work with Ken Ryan at West Virginia University and Steve Vardeman at Iowa State University.

Selecting an informative/discriminating multivariate response for inverse prediction

Edward Thomas, Albuquerque, New Mexico

Inverse prediction is important in a variety of scientific and engineering applications, such as to predict properties/characteristics of an object by using multiple measurements obtained from it. Inverse prediction can be accomplished by inverting parameterized forward models that relate the measurements (responses) to the properties/characteristics of interest. Sometimes forward models are computational/science based; but often, forward models are empirically based response surface models, obtained by using the results of controlled experimentation. For empirical models, it is important that the experiments provide a sound basis to develop accurate forward models in terms of the properties/characteristics (factors). While nature dictates the causal relationships between factors and responses, experimenters can control the complexity, accuracy, and precision of forward models constructed via selection of factors, factor levels, and the set of trials that are performed. Recognition of the uncertainty in the estimated forward models leads to an errors-in-variables approach for inverse prediction. The forward mod-

els (estimated by experiments or science based) can also be used to analyze how well candidate responses complement one another for inverse prediction over the range of the factor space of interest. One may find that some responses are complementary, redundant, or noninformative. Simple analysis and examples illustrate how an informative and discriminating subset of responses could be selected among candidates in cases where the number of responses that can be acquired during inverse prediction is limited by difficulty, expense, and/or availability of material.

Experiment Design for Big Data

Organizer: Xinwei Deng, Virginia Tech, and C. Devon Lin, Queen's University
Chair: Xinwei Deng, Virginia Tech

Projected support points - a new method for high-dimensional data reduction

Roshan J. Vengazhiyil, Georgia Tech.

This talk presents a novel method called projected support points (PSPs), which tackles the problem of compacting big data in high dimensions. This method has important applications to many practical problems in statistics and engineering, particularly when the available datais plentiful and high-dimensional, but the processing of such data is expensive due to computation or storage costs. In many applications, this data processing step requires functional evaluations at each data point, with the evaluated function exhibiting low-dimensional structure on the high-dimensional input space. The key novelty for PSPs is that it incorporates this low-dimensional structure within the data reduction procedure, by ensuring the data is well-reduced on low-dimensional projections of the data space. In this talk, we present a unifying framework for high-dimensional data reduction using PSPs, connecting important principles from experimental design, Quasi-Monte Carlo, and machine learning. We then introduce two algorithms for efficiently performing this reduction via difference-of-convex optimization. Finally, we illustrate the usefulness of PSPs in two real-world problems, the first from machine learning and the second from computer experiments. (based on joint work with Simon Mak.)

Discrepancy-based completely randomized design for A/B testing experiments

Yiou Li, DePaul University

A/B tests classically refer to the experiments conducted to estimate the treatment effect of a two-level control factor. The term is generalized to "A/B/n" when the control factor has more than two levels. The common way of conducting the A/B testing experiment is use completely randomized design that randomly assign treatment settings to the test subjects, who usually have some descriptive covariate variables. The purpose of this paper is to propose a criterion that measuring the discrepancy of the distributions of

covariate variables of different groups of the test subjects. According to the completely randomized design, each treatment level is to be randomly assigned to one of the groups. The discrepancy is based on the kernel density estimation. An optimization algorithm is developed to minimize this criterion in one-short for offline experiments and sequentially for online experiments. Both simulation and real data are used to show their performance.

Embracing experimental design thinking for large-scale statistical analysis

Peter Chien, University of Wisconsin-Madison

Experimental design is one of the fundamental contributions statisticians made to the society. It has led to major advances in science and engineering. Given the enormous success of experimental design in the past, it is imperative for statisticians to think how to best utilize this thinking in the Big Data era. Data with massive sample size and/or high dimensionality appear in observations, physical or simulation experiments, digital engineering, and many other sources. This work is motivated by the increasing need of subsampling a Hadamard matrix in statistical analysis methods as varied as compressed sensing, leverage score subsampling and matrix completion. The current solution to this problem, which randomly takes a subsample from a Hadamard matrix, suffers from high variability and large coherence. In this talk, we will discuss an experimental design framework to exploit substructures with low coherence in Hadamard matrices to accelerate large-scale statistical analysis. The effectiveness of the proposed framework will be corroborated by theoretical results and illustrated with examples from compressed sensing. Joint work with Youran Qi at the University of Wisconsin-Madison and Xu He at Chinese Academy of Science.

Data Science in New Mexico

Organizer: Emily Casleton, Los Alamos National Laboratory
Chair: Emily Casleton, Los Alamos National Laboratory

Overfitting in Bayesian model calibration of functional data under misspecified models

Lauren Hund, Sandia National Laboratory

Solving inverse problems by coupling experimental data with computational simulations is often a poorly identified problem, where multiple inputs produce equally valid solutions. When calibrating physical parameters using a misspecified model and in the presence of nuisance parameters, the "calibration solution" that best fits the observed data will typically not converge to the true values of the inputs. In this talk, we discuss strategies for physical parameter estimation using Bayesian model calibration when the

output is functional and the computer model is misspecified. Specifically, two alternatives to the standard Gaussian process model for discrepancy are considered. First, power-likelihood models are used to inflate the variance of the calibration parameters to account for model discrepancy. Second, the calibration parameters are allowed to vary over the functional output using temporally-varying covariate models. Discrepancy is gauged based on variability in the calibration parameters. These methods are applied to estimate dynamic material properties by coupling shock physics experimental data with computational predictions.

gibbSeq: a fully Bayesian multiple testing method for differential gene expression

Oleg Makhnin, New Mexico Tech.

gibbSeq is a fully Bayesian method for multiple testing based on hidden variables. It is primarily developed for the gene expression data, for example, RNA-seq. The method is based on lognormal distribution approximation for the RNA-seq read counts. It allows to directly estimate the FDR (false discovery rate) for the tests. The method also allows for direct testing of differential expression of gene sets, and may help account for lack of independence among the counts. From simulation studies, it performs really well compared to currently popular methods (edgeR, DEseq), when the data are indeed lognormal; the results are more mixed when the simulated data are Negative Binomial.

Using approximate Bayesian computation to infer evolutionary trees

James Degnan, University of New Mexico

Methods for inferring species trees from gene trees motivated by incomplete lineage sorting typically use either rooted gene trees to infer a rooted species tree, or use unrooted gene trees to infer an unrooted species tree, which is then typically rooted using one or more outgroups. Theoretically, however, it has been known since 2011 that it is possible to consistently infer the root of the species tree directly from unrooted gene trees without assuming an outgroup. Here, we use approximate Bayesian computation to infer the root of the species tree from unrooted gene trees assuming the multispecies coalescent model. It is hoped that this approach will be useful in cases where an appropriate outgroup is difficult to find and gene trees do not follow a molecular clock.

Celebrating 60 Years of Technometrics

Organizer: Daniel W. Apley, Northwestern University
Chair: Daniel W. Apley, Northwestern University

Replication or exploration? Sequential design for stochastic simulation experiments

Mickael Binois, Argonne National Laboratory

We investigate the merits of replication, and provide methods for optimal design (including replicates), with the goal of obtaining globally accurate emulation of noisy computer simulation experiments. We first show that replication can be beneficial from both design and computational perspectives, in the context of Gaussian process surrogate modeling. We then develop a lookahead based sequential design scheme that can determine if a new run should be at an existing input location (i.e., replicate) or at a new one (explore). When paired with a newly developed heteroskedastic Gaussian process model, our dynamic design scheme facilitates learning of signal and noise relationships which can vary throughout the input space. We show that it does so efficiently, on both computational and statistical grounds. In addition to illustrative synthetic examples, we demonstrate performance on two challenging real-data simulation experiments, from inventory management and epidemiology.

Permutation and grouping methods for sharpening Gaussian process approximations

Joseph Guiness, North Carolina State University

Vecchia's approximate likelihood for Gaussian process parameters depends on how the observations are ordered, which has been cited as a deficiency. This article takes the alternative standpoint that the ordering can be tuned to sharpen the approximations. Indeed, the first part of the paper includes a systematic study of how ordering affects the accuracy of Vecchia's approximation. We demonstrate the surprising result that random orderings can give dramatically sharper approximations than default coordinate-based orderings. Additional ordering schemes are described and analyzed numerically, including orderings capable of improving on random orderings. The second contribution of this paper is a new automatic method for grouping calculations of components of the approximation. The grouping methods simultaneously improve approximation accuracy and reduce computational burden. In common settings, reordering combined with grouping reduces Kullback-Leibler divergence from the target model by more than a factor of 60 compared to ungrouped approximations with default ordering. The claims are supported by theory and numerical results with comparisons to other approximations, including tapered covariances and stochastic partial differential equations. Computational details are provided, including the use of the approximations for prediction and conditional simulation. An application to space-time satellite data is presented as well.

Gaussian process modeling of a functional output with information from boundary and initial conditions and analytical approximations

Matthias Tan, City University of Hong Kong

A partial differential equation (PDE) models a physical quantity as a function of space and time. These models are often solved numerically with the finite element method and the computer output consists of values of the solution on a fine grid over the spatial and temporal domain. When the simulations are time-consuming, Gaussian process (GP) models can be used to approximate the relationship between the functional output and the computer inputs, which consists of boundary and initial conditions. The Dirichlet boundary and initial conditions give the functional output values on parts of the space-time domain boundary. Although this information can help improve prediction of the output, it has not been utilized to construct GP models. In addition, analytical solutions of the PDE derived by simplifying the PDE can often be obtained, which can help further improve performance of the GP model. This paper proposes a Karhunen-Love (KL) expansion based GP model that satisfies the Dirichlet boundary and initial conditions almost surely, and effectively uses information from analytical approximations to the PDE solution. Numerical examples demonstrate the improved prediction performance achieved by using these sources of prior information.

The Latest in Statistical Process Control and Signal Detection

Organizer: Anne Hansen-Musakwa, Tuft & Needle
Chair: Anne Hansen-Musakwa, Tuft & Needle

Repeated SPRT Charts for Monitoring INAR(1) Processes

Daniel R. Jeske, University of California, Riverside

Poisson integer valued autoregressive (INAR) models have been proposed for modeling correlated count data. Poisson lognormal (PLN) INAR models extend their use to overdispersed contexts. In this paper, we will propose the use of a repeated SPRT procedure to detect change in first-order INAR and PLN INAR models. We consider change in the mean, the autocorrelation parameter, and the overdispersion parameter. Simulation results show the repeated SPRT procedure performs favorably relative to previously proposed CUSUM procedures that are based on either the observations themselves or residuals of the observations from predicted values. A data set on invasive insect species is used to illustrate the repeated SPRT procedure.

Anomaly detection in multivariate and streaming data

Karl Pazdernik, Pacific Northwest National Laboratory

Algorithms capable of efficiently ingesting and describing streaming data have become

essential due to the growing popularity of social media platforms, wearable technology, and the Internet of things (IoT). A common problem found in all such domains is identifying when an observation is anomalous based on historic patterns. In this presentation, we focus on the specific statistical task of anomaly detection in multiple sources of streaming data. We contrast various forms of matrix and tensor decompositions and discuss the impact of utilizing the hierarchical structure of temporal and multivariate data. Finally, we show how this methodology can be used for anomaly detection in multiple forms of social media.

Process control using machine learning

Xin Guan, Intel Corporation

Traditionally, we use statistical process control methods to ensure process is predictable, stable and consistent and to detect anomaly in production. Control charts are often used to examine if a process is right on target and within a normal variation. However, most of control charts are one dimensional, that is, we only look at one parameter at a time. But as we know, we often need to examine multiple parameters in one process, especially for complicated processes. The relationships between the inspection outcome and each process parameter are usually non-linear and non-additive. Machine learning methods can help to put all important parameters in one model. By looking at the model prediction in a control chart, we are able to detect anomaly in tools and processes even before products go to inspection. It is extremely helpful if there is a latency in inspection as we can prevent yield loss upstream by prediction.

Statistical Machine Learning

Organizer: C. Devon Lin, Queen's University, and Matt Pratola, The Ohio State University
Chair: Xinwei Deng, Virginia Tech

Variable selection for mean and volatility

Rob McCulloch, Arizona State University

Dramatic advances in Bayesian modeling and computation have given us powerful tools for flexible fitting of high dimensional relationships. However, the flexibility and complexity of the modeling procedures comes at a price: we may have difficulty understanding what our models have found. In particular, we are often interested in finding a simple model that works well, with variable selection being an important special case. Recently Pratola et al. have developed a Bayesian approach to a model of the form Y = f(x) + s(x) Z where both the function f and s are learnt using ensembles of trees. This enables us to do variable selection for the mean (f) and the volatility (s) separately. Pratola et al. used simple measures of how often a variable is used in the ensembles. In this paper

we use the method of Carvlaho et al. to search for variable parsimonious approximators to the functions f and s.

Adaptively Pruned Random Forests for Modeling Means and Variances Simultaneously

Thomas M. Loughin, Simon Fraser University

Random Forests for regression are typically constructed using standard regression trees, which make splits that minimize squared error. Thus, the trees implicitly assume homoscedasticity, and have been found not to be robust against heteroscedasticity. This lack of robustness may be passed on to the forests. As an alternative, we use trees based on a heteroscedastic Gaussian likelihood as base learners in an ensemble much like a random forest. The new trees may split on means and/or variances, depending on which kind of split is best according to a specially derived information criterion. We show how to efficiently combine means and variances from different trees to produce estimates of means and variances for the full forest. We explore the performance of the new ensemble using both simulated and published data sets.

Survival prediction and model assessment via BART model averaging

Nick Henderson, Johns Hopkins University

In a wide variety of contexts, covariates which may be divided into distinct groups are used in constructing prognostic models of survival. To better utilize such groupings in high-dimensional patient covariates, we propose a flexible ensemble survival prediction method which relies on Bayesian model averaging of many tree-based accelerated failure time (AFT) models. The predictor space in each AFT model of our averaging scheme is derived from a specific clustering or grouping of the covariates, and for each of these, the corresponding AFT model consists of a Bayesian additive tree model for the regression function and a Gaussian error term. Additive regression trees allow for substantial flexibility in modeling the relationship between patient covariate and survival times within each AFT model, and Bayesian model averaging allows for more efficient exploration of the covariate space along with improved survival prediction and more interpretable model assessment. We explore the use of our method using a multi-site study of lung cancer survival and discuss its use in determining the relative importance of key clinical and genetic features.

Computer Experiments

Organizer: Shan Ba, Procter & Gamble
Chair: Brian Williams, Los Alamos National Laboratory

Practical heteroskedastic Gaussian process modeling for large simulation experiments

Robert B. Gramacy, Virginia Tech.

We present a unified view of likelihood based Gaussian progress regression for simulation experiments exhibiting input-dependent noise. Replication plays an important role in that context, however previous methods leveraging replicates have either ignored the computational savings that come from such design, or have short-cut full likelihood-based inference to remain tractable. Starting with homoskedastic processes, we show how multiple applications of a well-known Woodbury identity facilitate inference for all parameters under the likelihood (without approximation), bypassing the typical full-data sized calculations. We then borrow a latent-variable idea from machine learning to address heteroskedasticity, adapting it to work within the same thrifty inferential framework, thereby simultaneously leveraging the computational and statistical efficiency of designs with replication. The result is an inferential scheme that can be characterized as single objective function, complete with closed form derivatives, for rapid library-based optimization. Illustrations are provided, including real-world simulation experiments from manufacturing and the management of epidemics.

A latent variable approach for handling qualitative factors in Gaussian process modeling of computer experiments

Daniel W. Apley, Northwestern University

Computer simulations often involve both qualitative and numerical inputs. Existing Gaussian process (GP) methods for handling this mainly assume a different response surface for each combination of levels of the qualitative factors and relate them via a multiresponse cross-covariance matrix. We introduce a substantially different approach that maps each qualitative factor to an underlying numerical latent variable (LV), with the mapped value for each level estimated similarly to the covariance lengthscale parameters. This provides a parsimonious GP parameterization that treats qualitative factors the same as numerical variables and views them as effecting the response via similar mechanisms. This has strong physical justification, as the effects of a qualitative factor must always be due to some underlying numerical variables. Even when the underlying variables are many, sufficient dimension reduction arguments imply their effects can be represented by a low-dimensional LV. This is supported by superior predictive performance observed across a variety of examples. Moreover, the mapped LVs provide substantial insight into the nature and effects of the qualitative factors.

Constructing space filling designs with categorical factor and factor constraints

Bradley Jones, JMP Division/SAS

There are occasions when a deterministic computer code requires the user to include factors that take discrete unordered levels. In such a case, it is desirable that the design when considering only the continuous factors should have good space filling properties. It is also desirable that for each level of the categorical factor (or factors) the continuous factors should again have good space filling properties. We supply a design criterion and algorithmic approach that produces designs that provide both the above desirable properties. In addition, this methodology does not require the design region to be a hypercube. We provide examples demonstrating all these claims. This is joint work with Ryan Lekivetz.

Contributed Program

Topics in Spatial Statistics

Chair:

Comparison of alternative Kriging estimation methods including cross-validation

Long Wang, The Ohio State University

We proposed cross-validation methods for estimating Kriging model parameters motivated by the desire to address numerical issues in likelihood estimation and the goal of accurate prediction. We use numerical test problems and full factorial experimentation to study the effects of factors including the fitting objective (likelihood, cross-validation, and a convex approximation to likelihood), the experimental design (including designs with neighboring points), and the optimization method (including particle swarm heuristics). The results suggest that the proposed cross validation-based methods can provide potentially important advantages in prediction accuracy particularly in cases in which singularity issues make accurate likelihood estimation numerically difficult.

Anisotropic functional blind deconvolution with application to seismic inversion

Rida Benhaddou, Ohio University

We consider the problem of estimating a periodic two-dimensional function f based on observations from its noisy convolution when the convolution kernel itself is unknown but observations are available that are themselves contaminated with a noise. A preliminary thresholding procedure is applied to stabilize the inversion. We construct a wavelet hard-thresholding estimator of f, derive minimax lower bounds for the L^2-risk when f belongs to a Besov ball of mixed smoothness and show that the proposed estimator is adaptive and asymptotically near-optimal in a wide range of anisotropic Besov balls. In addition, it is demonstrated that the convergence rates are expressed as a maxima between two terms, taking into account both noise sources. The problem studied is

motivated by seismic inversions in geophysical explorations. The common practice in seismology is to recover layer structure separately for each profile $u_l, l = 1, 2, \cdots, M$, and then combine the derived estimates into a two-dimensional function. It is shown that the two-dimensional deconvolution out performs the M separate recoveries in many scenarios. In particular, we show that unless the function f is very smooth in the direction of the profiles, very spatially inhomogeneous along the other direction and the number of profiles is very limited, the functional deconvolution approach has a much better precision compared to a combination of M separate one-dimensional recoveries. To see how the proposed blind deconvolution performs in a finite sample setting a limited simulations study is carried out, which shows that with the right tuning the procedure can achieve a good precision.

Bayesian Perspectives

Chair:

Bayesian-type change-point detection in statistical process control

Michael Baron, American University

Bayesian multichannel change-point detection problem is studied for wide classes of stochastic processes arising in statistical process control. Under the loss function that penalizes for false alarms and detection delays, with increasing penalty after each missed change-point, we derive asymptotically pointwise optimal (APO) rules for sequential change-point detection. These APO rules are attractive because of their simple analytic form, straightforward computation, and weak assumptions. Real applications often involve multiple data streams, nuisance parameters, time-dependence, nonstationarity, and rather complex prior distributions. Proposed APO rules can operate under these conditions, achieving asymptotic optimality. An example of change-point detection in a multivariate autoregressive process is used to illustrate the proposed methods.

Computing tolerance bounds using Bayesian tools

Jose Ramirez, Amgen

Tolerance bounds and intervals are commonly used to set product specification limits, determine acceptance criteria for comparability studies between two processes or sites, and for setting validation acceptance criteria. Although common so]ware can calculate tolerance bounds (SAS, JMP, R, etc.) these are, mostly, for the normal distribution case. In this talk, we discuss how the Bayesian method described by Wolfinger (1998), Krishnamoorthy & Mathew (2009), and others can be used to derive reasonably accurate tolerance bounds not only for the normal case, but also for those situations where the normal distribution is not a good approximation to the data. We give examples of common situations in the biotechnology sciences where the lognormal, gamma, beta and

Poisson distribution are appropriate distributions, and show how easy it is to implement using PROC MCMC in SAS.

General Interest

Chair:

The art of teaching and communicating design of experiments to non-statisticians

Shari Kraber, Stat-Ease

Drawing from over 20 years of experience teaching design of experiments to non-statisticians and consulting with clients on DOE projects, Shari shares some lessons learned. She will provide several entertaining analogies used to help clients understand statistical concepts. (For engineers and chemists, golfing and helicopters can provide unique inspiration, thus overcoming those alpha and beta obstacles!) Discover some of the key teaching points that help non-statisticians overcome their fear of statistics and allow them to propagate good DOE practices. Communication being the key to success, Shari will finish up with tips and tricks for presenting experimental results compellingly to managers and clients.

Innovation, entrepreneurship, and textiles

Samenah Pourmojib, North Carolina State University

Almost every country, state, region and city now has innovation and entrepreneurship as a priority[1]. In the U.S. between one-third to one-half of economic growth can be attributed to innovation [2]. Although innovation and entrepreneurship have been studied extensively, very little research has focused on the textiles and apparel industry. Most of the business literature is weak analytically, the research methods are a mixture of quantitative and qualitative methods.

First, we are interested in the intersection of innovation and entrepreneurship. There are many definitions of innovation and success in the literature. What are Entrepreneurship success factors? What is innovation? Is there any significant relationship between these two among the data sets? Secondly, how textiles and apparel entrepreneurs are different from other industries? What are the predicting models on entrepreneurship on literature? Do these models fit to apparel and textiles entrepreneurs? Thirdly, how are government and non-government interventions driving innovation and entrepreneurship? How can we measure the impact of these interventions?

We defined success and innovation based on our literature review. Then using Nvivo, a qualitative data analysis package, to identify ?success factors? and ?innovation factors? frequency in our data set of personal interviews of more than 400 entrepreneurs, mostly from textiles and apparel industry firms. Then we are analyzing the frequencies to find regression parameters and also to understand the correlations that exist. We are also

reviewing all the mathematical entrepreneurship models in the literature. We will then test major hypotheses in our samples to evaluate if these models fit our samples from textiles and apparel entrepreneurs. We are testing the ?success factors? and ?innovation factors? through personal in-depth case studies with a sample of start-up companies in North Carolina supported by federal grants and matching state funds.

The reasons people become entrepreneurs seem to be similar across all industries. The success factors are also similar. True passion seems to be a universal success factor. There are differences in the success factors for entrepreneurial companies in textiles and apparel from companies in other industries that are being studied. Our review of over 300 companies in North Carolina supported by federal grants and matching state funds shows clearly that for many entrepreneurial companies these federal and state funds were absolutely critical for their success

SPC to Knock Your Socks Off

Chair:

Quality control charts not based on sigma limits

Mian Adnan, Indiana University Bloomington

Since Quality Control Charts are being used to decide the cut off points for the quality products in variety of industries, these charts should be as invariant as possible to the outliers. Non-robust method of finding limits (upper limit and lower limit along with its mathematical form) for both symmetric and asymmetric distribution have been suggested, considering simultaneously the appropriate measures of location and dispersion of a distribution. These limits are more representative and are relatively less affected by outlier(s).

Weak signal detection using SPC

Gejza Dohnal, Czech Technical University

Changes in the behavior of dynamic systems are detected based on changes in the monitored quantities or their characteristics. Detection usually takes place by monitoring the progress of the variable, and detecting the change at a time when the predetermined threshold is exceeded. This threshold is determined on the basis of the detection scheme requirements, in particular the probability of false alarms and the rate of detection of the actual change. In some cases, however, change does not come suddenly, but some "indicia" in system behavior can be observed, which may indicate future change. For example, an increasing frequency of outliers can result in a sudden permanent change in the signal. The occurrence of some "unusual" frequencies is often known as an imminent change. Increasing correlation indicates undesired process status and the like. Detection of these "subliminal" hints can often improve the characteristics of the detection scheme,

especially the rate of detection of the actual change. In this paper we will deal with the detection of weak signals in statistical process monitoring (SPC) using a control chart. The statistical process of monitoring and sequential detection scheme is con- sidered. Assuming slow changes of standard behavior of the monitored process, we check some irregularities, which become evident in time to indicate an up- coming change. In such case we can use the described methods to detect the change faster than using classical SPC methods, and with a low rate of false alarms.

One-Class peeling for outlier detection in high dimensions

Maria Weese, Miami University

Outlier detection is important for preprocessing data and/or for detecting anomalous observations. Numerous outlier detection methods have been proposed in the fields of statistics, machine learning, and data mining. Some outlier detection methods are based on distance measures and require estimation of the sample covariance matrix. These methods can be computationally infeasible for high dimensional data, especially when the dimension of the data exceeds the sample size. Other methods such as k-nearest neighbors, require adequate specification parameters to perform satisfactorily. In this paper we propose a flexible framework to detect multiple outliers that does not require covariance estimation, is well-suited to high-dimensional datasets with a high percentage of outliers, and is robust to parameter specification. We evaluate our framework using both synthetic and benchmark data sets, showing that it works well in high dimensions and performs better than benchmark methods, especially when there is a higher percentage of outliers. This presentation will present the method, simulation results and benchmark data performance results.

The structure of "ultimate intelligence" and a possible future for optimal experimental design

Theodore Allen, The Ohio State University

This article reviews the history of selected machine learning methods. A concept called "the structure of ultimate intelligence" is proposed that relates machine learning, statistics, and operations research (OR) methods. Ultimately intelligent methods include all the imaginable complications of decision such as parametric uncertainty (because of data insufficiency), model bias or miss-specification, multi-period decision-making, human objectives, and gaming. The benefits of the ultimate intelligence concepts include an ability to relate diverse methods such as topic models and reinforcement learning. Also, the future of machine learning, OR, and statistics perhaps inevitably involves all of these aspects. Further, the three decision problems related to experimental planning, analysis, and action taking are described. This establishes the relationship between optimal experimental design and machine learning and suggests opportunities for future research focusing on reinforcement learning for response surface contexts.

Screening and Surviving

Chair:

Optimal designs for Gamut models

William Heavlin, Google

Experiments exploring directions of steepest ascent (DSAs) are isomorphic to one-factor-at-a-time (1AAT) experiments ? they try runs along the DSA at varying distances, typically with intermediate factor levels (IFLs). A natural extension of the 1AAT DSA approach likewise encourages IFLs, incorporates the DSA, and includes also the factorDSA interactions. When combined with locally weighted scatterplot smoothing ("lowess"), this defines a class of varying-coefficient models called gamut models. Our main optimal design criterion is average (or integrated) mean square prediction error (I-optimality), which is relatively friendly to IFLs. We adapt this criterion to the lowess-based varying-coefficient gamut models and assess the properties of the resulting designs. Our motivating example involves a high-energy plasma reaction.

Augmenting definitive screening designs for estimating second-order models

Abigael Nachtsheim, Arizona State University

Jones and Nachtsheim (2011) introduced a class of three-level screening designs subsequently called definitive screening designs (DSDs). The structure of these designs results in the statistical independence of best estimators for main effects and two-factor interactions; the absence of complete confounding among two-factor interactions; and the ability to estimate all quadratic effects. Because quadratic effects can be estimated, DSDs can allow for the screening and optimization of a system to be performed in one step, but only when the number of terms found to be active during the analysis of the data from the initial DSD is less than roughly half the number of runs required by the DSD. Otherwise, estimation of second-order models requires augmentation of the DSD. In this paper we explore the construction of a series of augmented designs, moving from the starting DSD to designs capable of estimating the full second-order model. We use power calculations and model-discrimination criteria to identify the numbers of augmented runs necessary to effectively identify a specified number of active model effects.

Minimum contamination and beta-aberration criteria for screening quantitative factors

Po Yang, University of Manitoba

For quantitative factors, the minimum-aberration criterion is commonly used for examining the geometric isomorphism and searching for optimal designs. In this paper,

we investigate the connection between the minimum-aberration criterion and the minimum contamination criterion. Results reveal that in ranking designs by the two criteria, the optimal designs selected by them can be different. We provide statistical justifications showing that the minimum contamination criterion controls the expected total mean square error of the estimation and demonstrate that it is more powerful than the minimum -aberration criterion for identifying geometrically nonisomorphic designs.

An empirical saddlepoint approximation method for smoothing survival functions under interval-censoring

Manjari Dissanayake, Texas Tech University

Interval-censored data are a common type of incomplete data which occur in medical studies. A commonly used non-parametric procedure to estimate survival functions is through the Kaplan-Meier (KM) estimator. The distribution the KM estimator delivers is a discrete one with approximations to the distribution only at the observation times given. The proposed method is a non-parametric method to produce smooth KM survival functions using an empirical saddlepoint approximation. The resulting distribution is constructed by inverting the moment generating function (MGF) for the KM estimated discrete distribution and obtaining the cumulative distribution function (CDF). Simulation studies are conducted to demonstrate the performance of the method among competing parametric method and the semi-parametric spline-based method. Integrated square errors (ISE) are used to compare the obtained functions with the true underlying distributions.

Inference and Genomics

Chair:

Inference on location parameter under multivariate skew normal setting

Ziwei Ma, New Mexico State University

In this paper, the sampling distributions of multivariate skew normal distribution are studied. Confidence regions and hypothesis tests on the location parameter, , are obtained in two cases: the one with known scale parameter and shape parameter, and the other with known shape parameter. Also confidence regions for are constructed by the pivotal method, Inferential Models (IMs), and robust method. The hypothesis test on it is proceeded as well. For the case when the scale parameter is unknown, the skew-F distribution is defined and the inference on , based on the Hotelling's T^2, is obtained. For illustration of our main results, the graphs of confidence regions and curves of the power function for the bivariate skew normal populations are presented for each case.

Regularized regression and parameter selection methods in genomic data classification and disease prediction

Karel Kupka, TriloByte

In this contribution we describe a procedure for analyzing genomic data from patients with a known diagnosis and a control group. We aim at identifying significant gene expressions (variables) as effective predictors of the binary ?diagnosis? Bernoulli random variable. Clinical genomic data matrices were optimized by a stepwise information-cost-driven missing data deletion. Recently published stagewise regularized linear regression was used to select variables. Clinical genomic data was then modelled by a linear and Gaussian RBF Support Vector Machine classifier in the reduced-dimensional original feature space and the results are compared. We show that in most cases, a linear classifier provides better results than non-linear models thanks to their well-posedness and stability with respect to cross-validation.

Same-species contamination detection in next generation sequencing

Tao Jiang, North Carolina State University

Same-species contamination detection is an important quality control step in genetic data analysis in human genetic sequencing as samples might be contaminated by lab technicians or samples from other contributors. Compared with widely discussed cross-species contamination, same-species contamination is more challenging to detect and few methods have been published to address this issue. This article introduces a novel machine learning algorithm to detect same species contamination using support vector machines. Our approach uniquely detects such contamination using variant calling information stored in the variant call format (VCF) files (either DNA or RNA), and can differentiate between same species contamination and mixtures of tumor and normal cells.

In the first stage of our approach, a change-point detection method is used to identify copy number variations or copy number aberrations (CNVs or CNAs) for filtering prior to testing for contamination. Next, single nucleotide polymorphism (SNP) data is used to test for same species contamination using a support vector machine model. Based on the assumption that alternative allele frequencies in next generation sequencing follow the beta-binomial distribution, the deviation parameter ρ is estimated by maximum likelihood method. All features of a radial basis function (RBF) kernel support vector machine (SVM) are generated using training data from Q2 Solutions and publicly available source, and then applied in the test data to detect contamination. If training data is not available, a default RBF kernel SVM model is used.

We demonstrate the potential of our approach using simulation experiments with varying levels of contamination. Sequence data in the form of fastq files from two publicly available cell lines (NA12878 and NA10855 from the 1000Genomes project) was used to create synthetically contaminated samples with various levels of contamination. VCF

files were generated, and the power and false positive rate of our approach to detect same species contamination was evaluated. Our simulation experiments show that our method can detect levels of contamination as low as 5%, with reasonable false positive rates. We provide an R software implementation of our approach using the det_ct() function in the CTN R package.

Parameter inference in generalized population genetics models

Timothy Wallstrom, Los Alamos National Laboratory

We consider the problem of parameter estimation for a space of genealogical models, given genetic sequence data from a sample of individuals. When the offspring distribution falls off rapidly, such genealogies fall into a single universality class, known as the Kingman coalescent. Until recently, the Kingman coalescent and its variants were essentially the only models used in genealogical inference. In this sense, it played a role analogous to that often played by the Gaussian distribution.

Recently it has been discovered that when the offspring distribution falls off more slowly a family of coalescents emerge, which are qualitatively different from the Kingman coalescent. The relationship between the Kingman and generalized coalescents is closely connected to the relationship between the Gaussian and stable distributions. These generalized models appear to provide a better fit for many natural populations, but in order to exploit this new model space, it is necessary to be able to infer the correct model for a particular population.

Our problem is to estimate the parameter in a parameterized set of generalized coalescents, given sequence data. The statistical challenge is that (1) the likelihood is intractable, in the sense that it involves the integration over a space of latent variables, which in this case is the space of genealogical trees; (2) the probability of the observed data will be zero for nearly all randomly sampled trees. We show how the use of Godambe's unbiased estimating equations provides an efficient solution.

Cool Case Studies in Prediction

Chair:

Machine learning approaches for NASA mission support and financial systems data to predict employee travel expenses

Andy Ramlatchan, NASA Langley Research Center

Among all government agencies, NASA is at the forefront of developing machine learning approaches in the advancement of its mission. While this work has often been aimed at the research directorates, (e.g. aerospace engineering, atmospheric sciences, environmental simulations, crew state monitoring, aeroelasticity, etc.), there has been little

exploration into using predictive analytics for mission support and operations, particularly in financial modeling. Within that domain, a particular area of concern is around employee travel expenses. As a public entity with inflexible budgetary constraints, the use of travel funds and the return on investment (ROI) for such travel has been evaluated using modern data science methodologies to yield valuable business intelligence. This insight can have a tremendous impact on future planning, operations, and agency resource management.

This research investigation began by mapping all NASA business intelligence systems, with a particular interest in those data sources that contain travel related data. Some considerations were paid to systems that track employee activities such as conference attendance, papers published, or other research functions that resulted from a work related travel expense. Large datasets were then merged using common identifiers, after which data cleansing techniques were used to create uniformity among the various expressions of variables housed in multiple databases, data warehouses and other repositories. This created a single dataset for analysis that encompassed 170 variables from approximately 120,000 travel expenses in fiscal year 2017. Feature engineering was conducted to generate additional continuous variables for analysis, such as creating a variable for distance traveled based on the departure city and destination city.

After gaining a thorough understanding of the data and the domain, followed by data processing, exploratory data analysis was conducted using unsupervised machine learning methods. This began by extrapolating a correlation matrix for continuous variables, followed by multiple clustering methods to classify travel observations into groups based on the travel amount authorized (using K-means and partitioning around mediod). Travel records were grouped into classes regardless of the multitude of factors that can impact trip cost (e.g. per diem rates, destination, amount of days, distance traveled, etc.). Appending the cluster number to each row in the data set allowed this new variable to be used as another feature for modeling. A feed forward artificial neural network was developed and trained on the dataset to predict into what trip expense group future trips would be classified. An additional neural network was developed and trained to predict the amount that would be spent on a trip, if given certain inputs. The combination of these two models for classification and regression ultimately helped to predict how travel dollars are being spent, and how much can be expected in the future for individual trips or for large scale budget planning purposes. Additionally, they served as a framework for which additional questions could be investigated with minimal changes to the models outside of simple variable manipulation.

Convex clustering of generalized linear model with application on purchase likelihood prediction

Shuyu Chu, IBM T. J. Watson Research Center

Predicting the purchase likelihood of potential clients is essential in finding the optimal pricing strategy in modern business. However, the heterogeneity related to both clients and products results in very different purchase behaviors. Thus, building one

global model on all data is no longer appropriate. And fitting distinctive models under different data segments is of great need. Towards this aim, we proposed a convex clustering approach to performing data segmentation and model fitting simultaneously. The proposed method ensures data points with a common model structure are grouped into the same segment. Besides, an iteratively weighted least squares (IWLS) algorithm is developed to achieve better convergence properties in parameter estimation by linearizing the nonlinear objective functions. The performance of the proposed approach and its merits are illustrated by numerical examples and a case study with business data from a major IT service provider.

Predicting the future: Warranty & shelf-life

Chair:

Simple approach to calculate random effects model tolerance intervals to set release limits and shelf-life specification limits of pharmaceutical products

Richard Montes, Hospira

Tolerance intervals are used to statistically derive acceptance limits that drugs must conform to upon manufacture (release) and throughout shelf-life. The single measurement per lot in release data and repeated measurements per lot longitudinally for stability data have to be considered in the calculation. Methods for the one-way random effects model by Hoffman and Kringle (2005) [HK] for two-sided intervals and Hoffman (2010) [H] for one-sided limits are extended to a random intercepts, fixed slope model in this paper. The performance of HK and H were evaluated via simulation by varying the following factors: i) magnitude of stability trend over time, ii) sample size, iii) percentage of lot-to-lot contribution to total variation, iv) targeted proportion, and v) data inclusion. The performance metrics are average width (for two-sided) or average limit (for one-sided) and attained confidence level. HK and H maintained nominal confidence levels as originally developed, but H is too conservative (i.e., achieved confidence level exceeds the nominal level) in some situations. The HK method adapted for an attribute that changes over time performed comparably to the more computationally intensive generalized pivotal quantity (GPQ) and Bayesian posterior predictive (BayesPP) methods.

Warranty/Performance text exploration for modern reliability

Scott Wise, SAS Institute

With the increasing amount of warranty/performance data available, we have a golden opportunity to learn much more about our designs/products and come up with better reliability/maintainability models. However, this additional info often comes in the form of unstructured text from technicians, researchers, and customers. Inability to find

groupings and trends in this data can hide the real reliability impacting issues that our traditional data analysis on censored test data, warranty failure-time stamps, etc. won't be able to uncover. This session will seek to show how modern analytic so]ware can help provide a way to quickly and easily explore this unstructured text data to gain better insights into how our true reliability. We will cover the basics of how to visualize, group, analyze, model and ultimately find trends using warranty/performance unstructured text data. Industry examples will be shown using real warranty/performance data where text exploration was helpful in uncovering real reliability trends.

Fun With Computer Models

Chair:

Optimal calibration for computer model prediction with finite sample

Xiaowu Dai, University of Wisconsin-Madison

We consider a non-asymptotic frequentist framework for computer model prediction. This framework concerns two main issues: (1) many computer models are inadequate for physical systems and (2) only finite samples of physical observations are available for estimating model discrepancy and calibrating multivariate unknown parameters in computer models. We propose a method to achieve the optimal calibration and provide exact statistical guarantees in the sense that the predictive mean squared error is minimized with optimal calibration for any finite samples. We derive an equivalent formulation of optimal calibration which leads naturally to an iterative algorithm. The connection is built between the optimal calibration and the Bayesian calibration in Kennedy and O'Hagan [J. R. Stat. Soc. Ser. B. Stat. Methodol. 63 (2001) 425-464]. Numerical simulations and a real data example show that the proposed calibration outperforms the existing ones in terms of the prediction.

Functional nonlinear regression and registration using bayesian adaptive splines

Devin Francom, Los Alamos National Laboratory

We explore a class of nonparametric regression models for use with functional response, particularly when the functional response is misaligned. While traditional approaches to this problem would align the functional responses as a preprocessing step, our purpose is to build a model that can predict the functional response in its native, misaligned space. We approach this using Bayesian Adaptive Splines to model both warping functions (that register or align the data) and warped functional data in terms of a number of covariates. This framework is applied to emulation of a computer model with misaligned functional response.

Mixing it Up!

Chair:

Development of improved statistical methodology for eyewitness identification

Alice Liu, University of Virginia

Eyewitness lineups are used frequently in law enforcement and are crucial in thousands in criminal cases. However, law enforcement do not have the time and resources available to conduct the much-needed research for the development and validation of more reliable practices. Noted in the National Research Council (NRC) 2014 report, Identifying the Culprit: Assessing Eyewitness Identification, research in the effectiveness of law enforcement practices for eyewitness identification (EWI) procedures remains incomplete. In fact, in the EWI research field, there is a divide on how to analyze the data, collected in either a laboratory setting or real data from law enforcement agencies, and how to present the results. EWI errors have severe consequences, both in convicting the innocent and in freeing the true perpetrator to commit other crimes. A key aspect to this research is to produce more discerning statistical and analytical models for assessing EWI procedures, which take into account and maximize both positive and negative predictive values (PPV and NPV, respectively), as both types of errors are important considerations in policymaking. Currently, the statistical analyses only take into consideration PPV, but not NPV. Other statistical approaches that properly incorporate several covariates will also be considered.

Pointwise tolerance intervals for autoregressive models, with an application to hospital waiting lists

Kedai Cheng, University of Kentucky

Long waiting lists are a symbol of inefficiencies of hospital services. The dynamics of waiting lists are complex, especially when trying to understand how the lists grow due to the demand of a particular treatment relative to a hospital?s capacity. Understanding the uncertainty of forecasting growth/decline of waiting lists could help hospital managers with capacity planning. We address this uncertainty through the use of statistical tolerance intervals, which are intervals that contain a specified proportion of the sampled population at a given confidence level. Tolerance intervals are available for numerous settings, however, there are no standard or well-accepted approaches for autoregressive models. This talk fills that gap and introduces an approach for establishing pointwise tolerance intervals for classic autoregressive models. Some theoretical developments of tolerance intervals in this setting are discussed. A summary of simulation results will be presented, which demonstrates the good coverage properties of this approach. Finally, the method is applied to the monthly number of patients on hospital waiting lists in England.

A physics-specific change point detection method using torque signals in pipe tightening processes

Juan Du, Peking University

Change point detection in torque signals has been widely adopted for quality inspection during pipe tightening processes. Previous studies on the change point detection in this process generally focus on directly detecting the change points throughout torques without considering the underlying mechanism that generates various quasi-periodic nonlinear profiles, thereby introducing a series of false change points and increasing the risk of releasing defective pipes. To overcome this problem, we propose a novel change-point detection approach by fully considering the profile generating mechanism, and introduce a similarity weighted matrix with an integration of dynamic time warping and tightening physics, thereby reducing the chance of false change points. A weighted regression model is developed to determine the authentic change points by introducing the tightening process constraints. The performance of the proposed approach is demonstrated by both numerical and real case studies, and results show that the proposed method achieves a more effective detection power than the other existing methods in the pipe-tightening processes.

Purely sequential and two-stage bounded-length confidence interval estimation problems in Fisher's "Nile" example

Yan Zhuang, University of Connecticut

Fisher's "Nile" example is a classic which involves a bivariate random variable (X, Y) having a joint probability density function given by $f(x, y|\theta) = \exp(-\theta x - \theta^{-1}y),\ 0 < x,\ y < \infty$, where $\theta > 0$ is a single unknown parameter. We develop bounded-length confidence interval estimations for $P_\theta(X > a)$ with a preassigned confidence coefficient using both purely sequential and two-stage methodologies. We show: (i) Both methodologies enjoy asymptotic first-order efficiency and asymptotic consistency properties; (ii) Both methodologies enjoy second-order efficiency properties. After presenting substantial theoretical investigations, we have also implemented extensive sets of computer simulations to empirically validate the theoretical properties.

All About CUSUMs

Chair:

CUSUM and GLR charts for monitoring the scale parameter of right-censored Weibull lifetimes

Jaeheon Lee, Chung-Ang University

The lifetime is a key characteristic of product quality. It is best to obtain the lifetime data of all samples, but they are often censored due to time or expense limitations. In this paper, we propose a binomial cumulative sum (CUSUM) chart and a generalized likelihood ratio (GLR) chart for monitoring the mean of Type I right-censored Weibull lifetime data, for a fixed value of the Weibull shape parameter. We compare the performance of the proposed binomial CUSUM and GLR charts with CUSUM charts that studied previously using the steady-state average run length (ARL). The results show that the binomial CUSUM chart is better when the censoring rate is high and/or the sample size is small, and the GLR chart is better when the Weibull shape parameter is large and the censoring rate is not high.

Performance of risk-adjusted CUSUM chart under an incorrectly specified binary logistic regression model

Philipp Wittenberg, Helmut Schmidt University

Quality control charts used in a healthcare environment, for example, to monitor surgical performance are becoming more common. Risk-adjusted (RA) CUSUM charts, that utilize only raw risk scores like the Parsonnet score to assess the preoperative risk, may lead to a deterioration of the chart?s properties, in particular the false alarm behavior. Our approach considers the application of power transformations in the logistic regression model to improve the fit to the binary outcome data. From a list of alternatives, we derive an appropriate value for the power exponent δ. The average run length (ARL) to false alarm is calculated with the popular Markov chain approximation more efficiently by utilizing the Toeplitz structure of the transition matrix. A sensitivity analysis of the in-control ARL against the actually used value ? shows possible effects of incorrect choices of ? depending on the underlying patient mix. We show that these results can vary from robustness to severe effects (doubled number of false alarms).

More DOE Fun

Chair:

Multivariate Design of Experiments for Engineering Dimensional Analysis

Chris Nachtsheim, University of Minnesota

We consider the design of dimensional analysis experiments when there is more than a single response. We first give a brief overview of dimensional analysis experiments and the dimensional analysis (DA) procedure. The validity of the DA method for univariate responses was established by the Buckingham Π-Theorem in the early 20th century. We extend the theorem to the multivariate case and then develop basic criteria for multivariate design of DA experiments. Finally, we illustrate the construction of designs for

DA experiments for examples involving the design of a heat exchanger and a mechanical pump.

Designing experiments for dynamic responses

Rong Pan, Arizona State University

The dynamic response system to be discussed in this talk produces outputs that are a function of time or a function of an observable variable, what we refer to as the spectrum variable. Instead of a single response value, these systems generate response curves over the spectrum variable. The dynamic characteristics of these curves and their interactions with other experimental factors are of interest to experimenters, thus it requires new experimental design methods for exploring such system efficiently. For an experimenter, designing such experiments requires her to make decisions on (1) when or where to take response measurements along the spectrum variable and (2) how to choose the factor-level combination of controllable and noise variables. The first consideration is unique for such experiments, especially when the measurement cost is high. A good experimental design should require less experimental runs and less measurement frequency, while achieving a high statistical efficiency in parameter estimation or response prediction. In this talk, we present the B-spline model as a flexible nonparametric model for fitting dynamic response data. Based on the hierarchical modeling approach, mixed effects models can be formulated. We consider both experimental settings and sampling times simultaneously for D-optimal designs.

Exploring the Options

Narrow big data in streams and Kolmogorov complexity

Michal Cerny, University of Economics, Prague

In this talk, "Narrow Big Data" is formalized as a matrix $A \in \mathbb{R}^{n \times p}$, where n stands for the number of observations, p stands for dimension and n is significantly larger than p (formally, say that n is superpolynomial with respect to p). Assume that the size of available memory is bounded by a polynomial in p and that the rows of A are accessible on-line one-by-one. Once a data point is read into memory, it is dropped and is no more available. This is a natural computational model for Narrow Big Data supplied in a stream. The strong computational limitations?small memory and restricted access to data?reduce the available statistical toolbox for processing such datasets. For example, it is possible to prove that this computational model does not allow us to evaluate sample quantiles. Such negative proofs are based on Kolmogorov complexity: the main idea is to take A with high Kolmogorov complexity and show that if a quantity of interest were computable, then the entire dataset would be theoretically reproducible only from the information stored in the small memory. Conversely, there are also positive results: many statistical methods can be reformulated into this computational model with no

loss of generality. For example, two-step regressions, used e.g. in White?s Heteroscedasticity Test or Breusch-Godfrey?s Autocorrelation Test, can be efficiently evaluated under mild theoretical assumptions (such the length of lag in BG-test being $O(1)$). The main message of the talk is that, in principle, there are two classes of statistical procedures from the computational viewpoint: those computable in the described data model and those which are provably not computable. The first class is interesting because even if the dataset is huge, there is no need for data-reducing preprocessing steps (such as subsamplig or grouping). The second class of methods, for which negative proofs are available, justifies that the data reduction step is necessary if the method is to be applied, or that other methods computable in the restricted model need to be developed.

Design of experiments for bimatrix games with military and baseball applications

Olivia Hernandez, The Ohio State University

Attempting to develop desirable Nash equilibria policy distributions must often begin with a simulation or logistic regression model. For example, one may have an agent-based model of an aerial bombing involving an attacker and a defender and desire to determine a distribution on level settings for their decision factors. The situation is similar to robust engineering with player two corresponding to noise factors but, in bimatrix games, the problem must be symmetric. We are apparently the first to propose experimental designs and/or linear modeling for this problem and the first to formally explore the parametric stability of the associated equilibria. Compound and other orthogonal arrays are compared with respect to multiple efficiency measures. Both bombing and pitching policy development examples are described.

Poster Presentations

Range based limits for several measures of dispersion, comparisons and association for small sample

Mian Adnan, Indiana University Bloomington

Coefficient of Variation is widely being used in Quality Assurance Studies in Industries especially for determining the volatility. Variation of the relative-dispersion measure might not be instantly accessible. Besides, the available inequality (as for example, Bhattacharya inequality for explaining the variation of the measure of sample standard deviation) may difficult enough to understand for non-statisticians. In this circumstance, the range-based variation of the relative-dispersion measure can be simple enough for assessing the measure of relative-dispersion and its comparison. The present study ensembles the variation of coefficient of variation in terms of range for small sample. Smaller ranges having minimum upper limit of the sample relative-standard deviation and sample relative-mean deviation have been found. Attempts have also been taken to get the

range-based limits for measures of association viz correlation, etc.

Unsupervised machine learning for phase separation in lipid bilayers

Boian Alexandrov, Los Alamos National Laboratory

Phase separation in mixed lipid systems have been extensively studied both experimentally and theoretically because of its biological importance. However, the interpretation and analysis of Molecular Dynamics (MD) simulations representing temporal and spatial changes in such systems as well as the behavior and nature of the order parameter of phase separation is still a challenging task. Here, we present a new unsupervised machine learning method based on Nonnegative Matrix Factorization (NMF), called NMFk, that successfully extracts physically meaningful features from pre-selected datasets of trajectories generated by MD simulations of systems of mixed lipids. We use coarse-grained MD simulations of a membrane model composed of a ternary lipid mixture. Our results demonstrate that leveraging NMFk can (a) determine the role of different lipid molecules in a phase separation, (b) characterize the formation of nano-domains of lipids, (c) determine the timescales of interest and (d) extract physically meaningful features that uniquely describe the phase transition with broad implications.

Noisy edges and traits (NET): A differentially private data synthesis method for social network data

Claire Bowen, Los Alamos National Laboratory

For the last few decades, social network analysis (SNA) and research has grown in popularity; driven with the emergence of social media (e.g. Facebook and Twitter). SNA applications can range from following weather patterns to identify social contagion to tracking Twitter accounts to monitor mental health. Social network data is this diverse, because the data can be naturally represented as a relational network among entities. However, privacy protection is imperative in social network data since this data often contains sensitive individual information such as medical information. DIfferentially Private Data Synthesis (DIPS) techniques produce synthetic data and pseudo individual records in the differential privacy setting, a robust concept for privacy protection. We propose a new DIPS approach called Noisy Edges and Traits (NET), which protects both the nodal attribute and edge information. The application of NET on three case studies (business email exchanges and social media data sets) suggests that NET is easy to implement and is a valid approach to preserve privacy protection for social network data.

A data analytic framework for physical fatigue management using wearable sensors

Ying-Ju Tessa Chen, University of Dayton

Fatigue is one of the most important safety concerns in manufacturing industry which has motivated many researchers to investigate data mining methods to develop a model to detect the fatigue state before it happened. In this research work, we aim to answer the following questions:

1. How to predict fatigued and not-fatigued states using the accelerometer and heart rate sensors by applying machine learning and statistical models?
2. How will be the performance of the model if we limit the number of sensors?
3. How to define the criteria for the important features of the fatigued and not-fatigued states?

We discuss how human physical fatigue can be classified using analytical models. Best subset selection method via bootstrapping approach were used to filter and identify the important features for classification. Experimental results demonstrate the effectiveness of the proposed Ensemble methods. In addition, the sensor combination comparisons are conducted in order to find the most informative sensor in the model. Identification of the predictors presents important retrospective findings, which can be the basis for a prospective study. we can use the torso and heart rate sensors together for both of the MMH and SI tasks which would results in 85.1% and 86% average classification accuracy for the these tasks respectively. This sensor reduction demonstrates the financial benefits of using our analytical model.

Nonparametric bagged estimators for binary response data

David Collins, Los Alamos National Laboratory

Binary (0/1) response data, encountered in applications ranging from market research to toxicology, is typically analyzed using logistic or probit regression. These methods imply a latent threshold variable subject to a parametric model (logistic or Gaussian, respectively), and can produce misleading results if the true threshold model does not match the assumption. Using simulations and real data from explosives sensitivity testing, we present two bagged estimators of the response distribution, based on averaging the results of hundreds of bootstrap replications. Each replication uses either a binary classification tree (resulting in a random forest estimate) or a k?nearest neighbor algorithm. Estimators are optimized using Brier scoring (mean squared prediction error). We show convergence to the true model for large samples, and illustrate how a more accurate estimate results when the true model is far from logistic or Gaussian.

Exploring and investigating contributing factors of injury severity of drivers of emergency vehicles in Ohio

Peter Hovey, University of Dayton

The purpose of this study was to examine the contributing factors and characteristics associated with fatality and injuries sustained by drivers of emergency vehicles (EVs) involved in traffic crashes in the state of Ohio. Emergence vehicle drivers considered in this study include drivers of firetrucks, ambulances (emergency medical services), and law enforcement vehicles. Some few research efforts recently conducted using Ohio?s crash data have shown that crashes involving emergency vehicles tend to have increased crash and injury severity levels. The current study investigated the injury risk factors of crashes involving EVs by using Ohio crash data for 2011-2015. The data were analyzed with both logistic regression and a classification tree to identify factors associated with injury severity. Results of the two methods are compared.

Input-dependent calibration for large computer experiments

Jiangeng Huang, Virginia Tech

With remarkable advances in computing power, complex physical systems can now be simulated to high accuracy using computer experiments. A limited amount observations from field experiments can provide valuable information to improve the fidelity of computer experiments. Kennedy and O'Hagan (2001) provides a pretty flexible Bayesian calibration framework, assuming the best fitting value of the calibration parameter is fixed across the entire input space. Motivated by a real large computer experiment from General Electronic, we aim to develop new input-dependent calibration methodology for high dimensional computer experiments with multiple outputs, especially when there is evidence that calibration parameters are dependent on the value of controllable input variables. We develop a new input-dependent multi-output calibration method that tangentially addresses several additional challenges arising in a motivating real-data application provided by collaborators at GE Oil & Gas: high- dimensional input spaces, "missing" simulation output due to numerical instabilities, and local modeling to address the inherent non-stationarity in the simulation dynamics.

Marginal probabilities for conditionally specified logistic regression: An application with public health data

Curtis Miller, University of New Mexico

Conditionally specified logistic regression (CSLR) was introduced by Joe & Liu in 1996. CSLR models outcomes $Y_1, \ldots, Y_p$ with potential covariates $X_1, \ldots, X_m$. In CSLR, we not only find the association of response Y_i to covariate X_j (regression coefficients), but the relation of responses Y_i and Y_j (interaction terms). Y_j becomes a predictor term in

the logistic regression for Y_i: that is, we have a logistic regression for Y_i given other Y's and the covariates X. It is easy to calculate $P(Y_i = 1|Y_j, j \neq i, X_j, j = 1, \ldots, m)$. For modeling, we would like the marginal probabilities $P(Y_i = 1|X_j, 1 = j, \ldots, m)$. We show how marginal probabilities may be derived. This method is applied to a data set from the Navajo Birth Cohort Study, and results compared to those of two other methods. This approach was developed to deal with a complex problem in environmental public health; it may be applicable to any situation, such as may arise in monitoring industrial processes, where there could be multiple binary responses representing different phenomena that may be interrelated.

The construction of ϵ-bad covering arrays

Kevin Quinlan, Penn State University

Covering Arrays are commonly used in software testing since testing all possible combinations is often impossible. A t-covering array covers all factor level combinations for the projection of the design into any t factors. In some high cost scenarios testing 100% of t factor combinations is infeasible. In this work, the assumption that all factor level combinations for any projection must be covered is relaxed. An ϵ-bad Covering Arrays covers only $(1 - \epsilon)\%$ of the factor levels combinations required for a t-covering array across the entire design. Some theoretical bounds for constructions of this type exist, but an explicit construction method is lacking. This work presents the first construction method for this type of array and results in a higher coverage than a given known result. The explicit construction extends to an infinite number of factors, any fixed number of factor levels, and any strength t, where no systematic method had previously existed. The calculation of the exact value of ϵ is detailed as well as the limiting values when $l \to \infty$. In situations where the cost of running additional experiments is high relative to the cost of missing an error, or initial results are needed quickly these designs would be preferred over traditional t-covering arrays due to the high coverage obtained in a small number of runs. Follow up experiments can be done using a known construction method to obtain full t-coverage starting from an ϵ-bad array for some cases. It is shown how intermediate steps of this construction can be used as Partial Covering Arrays. Finally, a case study shows the practical use of designs of this type in a hardware security testing application.

A Bayesian nonparametric approach to multistate models

Richard Warr, Brigham Young University

First passage distributions in multistate models are used to answer important research questions. A popular multistate model is the semi-Markov process (SMP). An SMP has defined states, and often modelers choose parametric distributions to represent the waiting time from transitioning from one state to another. However in some situations, modelers may not want to make parametric assumptions about transition distributions.

Also, in processes with many states it is challenging to choose an appropriate distributions for each possible transition. To address these and other situations, we propose a Bayesian nonparametric method to model the waiting time between states. This method assumes a Dirichlet Process (DP) prior on each state to state waiting time distribution. The DPs are then updated with the collected data. The first passage distribution from one state to another is then a convolution and mixture of these posterior processes. Our method computes and makes inference on these first passage processes in semi-Markov models.

Resiliency assessment for automated mobility

Joanne Wendelberger, Los Alamos National Laboratory

Extreme environmental or demand scenarios may severely impact automated vehicles and associated infrastructure in ways that are not yet well understood. With the anticipated growth of automated, on-demand mobility, there is a need for modeling and analysis capability to understand the behavior of mobility systems and how they will respond under abnormal conditions. As automated vehicles begin to replace traditional vehicles, new usage patterns and societal dependencies are likely to emerge, making these systems an increasingly critical part of the transportation infrastructure. Like other transportation systems, a variety of types of disruptions can impact automated, on-demand mobility systems, including severe weather events, fires, large special events, power outages, gas leaks, and human generated activities such as protests or attacks. This raises the question of what will happen to these systems when their flexible, demand-driven architectures are pushed to their limits. Our approach combines statistical modeling of distributions and extremes with socio-technological analysis to identify and examine potential impacts of extreme conditions. Knowledge gained from modeling disruptions can be used to assess and improve the resiliency of systems under extreme circumstances. Many analyses rely on modeling of average behavior, but to understand extremes, it is important to look at the variability and tail behavior of the distributions of interest.

Parametric analysis for the variability in high-performance computing systems using mixture distributions

Li Xu, Virginia Tech

Fatigue is one of the most important safety concerns in manufacturing industry which has motivated many researchers to investigate data mining methods to develop a model to detect the fatigue state before it happened. In this research work, we aim to answer the following questions:

1. How to predict fatigued and not-fatigued states using the accelerometer and heart rate sensors by applying machine learning and statistical models?

2. How will be the performance of the model if we limit the number of sensors?

3. How to define the criteria for the important features of the fatigued and not-fatigued states?

We discuss how human physical fatigue can be classified using analytical models. Best subset selection method via bootstrapping approach were used to filter and identify the important features for classification. Experimental results demonstrate the effectiveness of the proposed Ensemble methods. In addition, the sensor combination comparisons are conducted in order to find the most informative sensor in the model. Identification of the predictors presents important retrospective findings, which can be the basis for a prospective study. we can use the torso and heart rate sensors together for both of the MMH and SI tasks which would results in 85.1% and 86% average classification accuracy for the these tasks respectively. This sensor reduction demonstrates the financial benefits of using our analytical model.